sciendo

DOI: 10.2478/cdem-2020-0002 CHEM DIDACT ECOL METROL. 2020;25(1-2):35-56

Vasil SIMEONOV[1]

BASIC MULTIVARIATE STATISTICAL METHODS FOR ENVIRONMENTAL MONITORING DATA MINING: INTRODUCTORY COURSE FOR MASTER STUDENTS

Abstract: The present introductory course of lectures summarizes the principles and algorithms of several widely used multivariate statistical methods: cluster analysis, principal components analysis, principal components regression, N-way principal components analysis, partial least squares regression and self-organizing maps with respect to their possible application in intelligent analysis, classification, modelling and interpretation to environmental monitoring data. The target group of possible users is master program students (environmental chemistry, analytical chemistry, environmental modelling and risk assessment etc.).

Keywords: chemometrics, environmetrics, exploratory data analysis (EDA), master students course

Introduction

The scientific discipline chemometrics has won a high level of prestige in the recent three decades. Although many scientists state that there is no specific focus of the chemometrics since it is split between many different fields of application among them analytical chemistry, organic chemistry, theoretical and computational chemistry, environmental chemistry, chemical technology, geochemistry, hydrochemistry etc. Data analysis is needed everywhere but the approach to the different goals of the chemometric study could be radically different. That is why different definitions of the chemometric strategy and ultimate goals could be found. It is our conviction that chemometrics as applied to the problems of the environmental chemistry could be named environmetrics and its main goals are classification, modelling and interpretation of environmental monitoring data sets.

In many environmental studies different patterns have to be determined and interpreted [1-5] just to mention few out of many examples. For instance, could monitoring data be used to classify various sampling sites along river catchments? If sets of different sites come into different patterns, it seems possible to optimize the number of sampling point in the monitoring net using the patterns formed as information source for the river water quality instead of using each one of the separate sampling points. Is it possible to identify latent factors responsible for the data structure of the monitoring net consisting of mutually correlated chemical parameters of the water quality? If the answer is positive, then a model of the environmental system could be constructed informing on the impact of different

[1] Faculty of Chemistry and Pharmacy, University of Sofia “St. Kl. Okhridski”, 1164 Sofia, J. Bourchier Blvd. 1, Bulgaria, email: vsimeonov@chem.uni-sofia.bg

natural or anthropogenic factors on quality of the system. Thus, the tasks for risk assessment and risk management become more real and applicable.

Exploratory data analysis (EDA) is based mainly on the multivariate statistical methods called principal components analysis (PCA) and factor analysis (FA). Using these methods one could visualize a multivariate data set, to detect relationships in a reduced coordinate space both between object of interest and variable characterizing the objects.

Another group of methods are often named unsupervised pattern recognition. Cluster analysis is one of the mostly applied from this group. Its goals are to construct a scheme of similarities between different features, called dendrogram, in which more closely related objects are closer to each other. The main branches of the tree-like scheme represent deviations from the similarities. Unsupervised pattern recognition differs from exploratory data analysis in that the aim of the method is to detect similarities whereas using EDA there no particular insight whether at all or how many groups will be detected. Very often it is said that the unsupervised pattern recognition is a spontaneous classification method without any preliminary training of the data set to follow a specific requirement to classify the data into a preliminary chosen number of similarity groups.

There are a large number of methods for supervised pattern recognition, mostly dedicated to classification problems. The chemometricians have created many discriminant methods where a training set of knowing groupings has to be available in advance. The question then is to find to which preliminary formed class (group) belongs an unknown sample (object).

Basic environmetric methods

In this lecture some basic principles of the mostly used environmetric methods will be presented.

Principal components analysis (PCA)

Principal component analysis (PCA) seems to be the most widespread multivariate chemometric technique and is a typical display method (also known as *eigenvector analysis, eigenvector decomposition* or *Karhunen-Loéve expansion*) [6]. It enables revealing the "hidden" structure of the data set and helps to explain the influence of latent factors on the data distribution. PCA is done on covariance matrix when the data are centered or on correlation matrix when the data are standardized. PCA transforms the original data matrix into a product of two matrices, one of which contains the information about the objects or cases (e.g. sampling sites from a monitoring net) and the other about the features (e.g. chemical or physicochemical parameters determining the quality of the environmental system in concentration units).

If the input data set is represented as a matrix containing *m* number of rows and *n* number of columns the main goal of PCA is to decompose the initial data set into matrix of the score vectors and matrix of the loadings vectors:

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1s} \\ a_{21} & \dots & a_{2s} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{ms} \end{pmatrix} \times \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1n} \\ f_{21} & f_{22} & \dots & f_{2n} \\ \dots & \dots & \dots & \dots \\ f_{s1} & f_{s2} & \dots & f_{sn} \end{pmatrix}$$

Thus, the matrix characterizing objects contains the scores (understood as projection) of objects on principal components (PCs). The other one, characterizing features is a square matrix and contains the set of eigenvectors (understood as weights, in PCA terminology called "loadings") of the original features in each PC. In matrix terms, this can be expressed as:

$$X = S \cdot L + E \tag{1}$$

where X is the original data matrix (features as columns, cases as rows), S is a scores matrix (has as many rows as the original data matrix), L is a loadings matrix (has as many columns as the original data matrix), E is an error matrix.

The number of columns in the matrix S equals the number of rows in the matrix L. It is possible to calculate scores and loadings matrices as large as desired, provided the "common" dimension is no larger than the smaller dimension of the original data matrix, and corresponds to the number of PCs that are calculated. Each scores matrix consists of a series of column vectors, and each loadings matrix a series of row vectors. Many authors use s_a and l_a notation to express these vectors, where a is the number of the PC. The matrices S and L are composed of several such vectors, one of each PC. The first scores vector and first loadings vector are often called the eigenvectors of the first PC. Each successive component is characterized by a pair of eigenvectors. Using f eigenvectors in one dimension, where f is smaller than, or equal to the rank of the data, f PCs can be obtained. Usually, a small number of PCs is needed to represent most of the information in the data. The minor PCs which explain little of the data structure can be eliminated, thus simplifying the analysis. Also, these minor PCs contain most of the random error, so eliminating them tends to remove extraneous variability from the analysis. In the ecosystems monitoring studies PCA and related multivariate techniques are often applied to determine the possible influence and contribution of natural and anthropogenic factors in data structuring.

The decomposition of the input matrix is achieved by transformation of the major axes of the correlation matrix R. The problem solution with the eigenvectors values:

$$R \cdot e_1 = \lambda_1 \cdot e_1 \tag{2}$$

$$R \cdot e_2 = \lambda_2 \cdot e_2 \tag{3}$$

where: e - eigenvectors, λ - eigenvalues,
leads to the following determinant:

$$|R - \lambda I| = 0 \tag{4}$$

Its solution brings couples of eigenvectors and eigenvalues with following properties:

- the eigenvalues are measure for explanation by the identified latent factors total variation $\left(s^2_{total}\right)$ of the correlation matrix R,
- the eigenvalues are ordered in decreasing rank:

$$\lambda_1 > \lambda_2 > \ldots \lambda_m,$$

- the sum of all eigenvalues is equal to the number of the features (variables) m,
- the eigenvectors are orthogonal to each other,
- the eigenvectors contain the non-normalized coefficients of the factor loadings matrix A.

The eigenvalues are normalized by division by $\sqrt{\lambda_j}$. This normalization leads to the values of the factor loadings in matrix *L*. These factor loadings have sense of statistical weights (they take values within the interval [0, 1]) and indicate the weight of the starting features in the newly formed latent factor.

The new characteristics (latent factors) do not correlate each other and along with the eigenvalues explain the total variation of the input data set. Since

$$\sum_{j=1}^{m} \lambda_j = m \tag{5}$$

the part of the variation explained by factor *j* is:

$$\frac{s_j^2}{s_{total}^2} = \frac{\lambda_j}{m} \tag{6}$$

Some important features of PCA could be summarized as follows. The principal components axes (the axes of the hidden variables) are orthogonal to each other. Most of the variance of the data is contained in the first principal component. In the second component there is more information than in the third one etc. For interpretation of the projected data both the score and the loading vectors are plotted. In the score plots, the grouping of objects can be recognized. A loading plot reveals the importance of the individual variables with respect to the principal component model.

A very important task in PCA is the estimating the number of principal components necessary for a particular PC model. Several criteria exist in determining the number of components in the PCA model:

(i) percentage of explained variance,
(ii) eigenvalue - one criterion (Kaiser criterion),
(iii) Scree - test,
(iv) cross validation.

The percentage of explained variance is applied in sense of a heuristic criterion. It can be used if enough experience is gained by analyzing similar data sets. If all possible principal components are used in the model the variance can be explained by 100 %. Usually, a fixed percentage of explained variance is specified, e.g. 80 %. In environmental studies even 75 % of explained variance is a satisfactory measure for the adequateness of the PCA model chosen.

The eigenvalue - one criterion (Kaiser criterion) is based on the fact that the average eigenvalue of autoscaled data is just one. In this case only eigenvalues greater than 1 are considered important.

The Scree - test is based on the phenomena that the residual variance levels off when the proper number of principal components is reached. Visually the residuals or more often the eigenvalues are plotted against the number of latent factors in a Scree plot. The principal component number is then derived from the leveling-off in the plot.

The forth approach of deciding on the number of principal components uses the following idea. In the simplest case, every object of the input matrix *X* is removed (*leave-one-out method*) from the data set once and a model with the remaining data is computed. Then the removed data are predicted by the use of the PCA model and the sum of the square root of residuals over all removed objects is calculated. In case of

large data sets, the leave-one-out method can be replaced by leaving out a whole group of objects.

Interpretation of the results of PCA is usually carried out by visualization of the component scores and loadings. In the score plot, the linear projection of objects is found, representing the main part of the total variance of the data (in the plot PC1 vs. PC2). Other projection plots are also available (e.g. PC1 vs. PC3 or PC2 vs. PC3) but they represent less percentage of explained total variance of the system in consideration. Correlation and importance of feature variables is to be decided from the factor loadings plots.

In environmental studies the interpretation of the monitoring data set using PCA is concentrated mainly on visualization of the patterns of objects (e.g. sampling locations, sampling periods, sampling dimensions or fractionalization etc.) in the factor scores plots, where the factor scores are actually the new coordinates of the objects in the reduced factor space. Another important option is consideration of the factor loadings plots and tables where the relationship between the features characterizing the objects could be observed and interpreted.

As already mentioned above, the elements of the matrix L mark the participation of each input feature (variable) in the formation in a certain latent factor. Variables having low values of the factor loadings negligibly influence the factor since those with high or negative loadings have a serious impact in the formation of the latent factor. Indeed, the factor loadings values of the input features are important indication for its nature. In environmental studies the latent factor identification (very often each latent factor is conditionally named in order to stress its nature and origin) makes it possible to detect the real natural or anthropogenic factors influencing the ecosystem and controlling its quality and equilibrium.

Very often the latent factors structure is not well defined. The participation of many input features into one latent factor hinders its interpretation. In order to simplify the latent factor structure a rotation of the coordinate system of the factors is applied. In this case the location of the objects does not change. The final result is the reduction of the number of the "average" factor loadings values (their absolute value counts) and respective increase of those close to zero average value and those whose value is high. The simplified factor structure could be achieved by orthogonal or non-orthogonal transformation. In many environmental studies this new structure describes quite satisfactory the impact of the natural or anthropogenic real factors affecting the ecosystem (Varimax rotation mode).

The application of PCA in an environmental study could solve problems related to:

- detection and determination of the structure of the real environmental factors controlling the quality and the state of a certain environmental compartment
- clarification and visualization of the structure of the object of interest
- modelling of the system in consideration keeping the mind the impact of the different identified by PCA factors.

Thus, the principal components analysis is a well-known projection method which makes it possible to assess the mutual relationships hidden in the monitoring data set. At the same time PCA is also a useful data modelling approach. The opportunity to construct a linear combination of the original variables in the data set (columns in the input matrix) being a better description of the features characterizing the objects (the rows in the input matrix) leads to kind of abstract measurements in the system (the dimension of the input data does not play any role; the latent factors turn to be better descriptors of the system). The final output brings more information about the system than the traditional

chemical, physical or biological indicators. The abstract variables (latent factors, principal components) require specific interpretation since a small number of them describe a large part of the system variation and in this way reduce its dimensionality and allow its proper visualization.

Cluster analysis

Exploratory data analysis (just like the PCA) is used dominantly to determine general relationships between data [7]. Very often the data treatment procedure has to respond to more complex questions about possible similarity between monitoring results or between sampling sites, i.e. question about formation of groups within the data set. Cluster analysis (CA) is a well developed strategy to determine relationships between different objects of observation characterized by various features or, *vice versa*, between the features describing the different objects. In such cases the unsupervised pattern recognition is the tool to solve the problem and cluster analysis along with some other methods is used to group different objects. CA enables objects stepwise aggregation according to the similarity of their features. As a result hierarchically or non-hierarchically ordered clusters are formed. A single cluster describes a group of objects that are more similar to each other than to objects outside the group. Similarity understood in the term of CA measures how alike two cases are. While the term similarity has not unique definitions, it is common to refer to all similarity measures as "distance in multi-features space" measures since the same function is served. A similarity between two objects *i* and *i'* is a distance if:

$$(D_{i'i} = D_{ii'}) \leq 0 \text{ where } D_{ii'} = 1 \text{ if } x_i = x_{i'} \tag{7}$$

where x_i and $x_{i'}$ are the row-vectors of the data table X with the features measurements describing objects *i* and *i'*. When two or more features are used to define their similarity, the one with the largest magnitude dominates. This is why primary standardization of features becomes necessary. There are a variety of different measures of inter-cases distances and inter-cluster similarities and distances to use as criteria when merging nearest clusters into broader groups or when considering the relation of an object to a cluster. A few most popular ways of determining how similar interval measured objects are to each other are as follows:

1. *Euclidean distance* - the distance between two objects x_i and $x_{i'}$ is defined by equation (2) where *j* presents repetition of measurements.

$$d_{x_i x_{i'}} = \sqrt{\sum_{j=1}^{J} (x_{ij} - x_{i'j})^2} \tag{8}$$

2. *Squared Euclidean distance* - removes the sign and places greater emphasis on objects further apart, thus increasing the effect of outliers.

$$d_{(x_i, x_{i'})} = \sum_{j=1}^{J} (x_{ij} - x_{i'j})^2 \tag{9}$$

3. *Manhattan distance (city-block distance, block distance)* is the average absolute difference across the two or more dimensions which are used to define distance. The Manhattan distance is defined slightly differently to the Euclidean distance. Except for some specific cases when Manhattan distance is equal to Euclidean distance, it is always higher than Euclidean distance.

$$d_{(x_i,x_{i'})} = \sum_{j=1}^{J} \left| x_{ij} - x_{i'j} \right| \tag{10}$$

4. *Chebychev distance* is the maximum absolute difference between a pair of cases on any one of the two or more dimensions (features) which are being used to define distance. Pairs will be defined as different according to their difference on a single dimension, ignoring their similarity on the remaining dimensions.

$$d_{(x_i,x_{i'})} = \max \left| x_i - x_{i'} \right| \tag{11}$$

5. *Mahalanobis distance* takes into account that some features may be correlated and so defines roughly the same object's properties (*C* is the variance-covariance matrix of the features).

$$d_{ii'} = \sqrt{\left(x_i - x_{i'}\right) \cdot C^{-1} \left(x_i - x_{i'}\right)'} \tag{12}$$

6. *Minkowski distance* should be applied if the object weight is increasing related to the dimensions in each compared objects and indicates the lowest similarity.

$$d_{(x_i,x_{i'})} = \sum_{j=1}^{J} \sqrt[r]{(x_i - x_{i'})^p} \tag{13}$$

7. *Pearson correlation* is based on correlation coefficient. Since for Pearson correlation, high negative as well as high positive values indicate similarity, the researchers usually select absolute values.

There are several other related distance measures (*weighted Euclidean distance*, *standardized Euclidean distance*, *cosine*, *customized*, etc.) but usually specific reasons are required if a very sophisticated distance measure is to be applied.

In case of CA one task is related with determination of similarity between measured objects, but equally important is to define how objects or clusters are combined at each step of similarity assessment procedure. One possibility for clustering objects is their hierarchical aggregation. In this case the objects are combined according to their distances from or similarities to each other. Within hierarchical aggregation agglomerative and divisive methods can be distinguished. Divisive clustering is based on splitting the whole set of objects into individual clusters, while in case of more frequently used agglomerative clustering one starts with single objects and gradually merges them in broader groups. Usually some objects create one broader group, while rest of them creates the other. As in case of distance measure various algorithms (linkage techniques) are available to decide on the number of clusters. They result in slightly different clustering pattern. A few most popular linkage algorithms are:

1. *Nearest neighbor (single linkage)* - the distance between two clusters is the distance between their closest neighboring objects, in other words the similarity of the new group from all other groups is given by the highest similarity of either of the original objects to each other object.

$$d_{mj} = \frac{d_{ij} + d_{i'j}}{2} - \frac{\left| d_{ij} - d_{i'j} \right|}{2} = \min(d_{ij}, d_{i'j}) \tag{14}$$

where: *m* - new object or cluster, *i*', *i*, *j* - clustered before objects.

This algorithm performs well when the plotted clusters are elongated or chain-like, moreover the sizes of the clusters and their weights are presumed to be equal.

2. *Furthest neighbor (complete linkage)* - the distance between two clusters is the distance between their furthest member objects. Furthest neighbor algorithm of linkage refers only to the calculation of similarity measures after new clusters are formed, and the two clusters (or objects) with highest similarity are always joined first.

$$d_{mj} = \frac{d_{ij} + d_{i'j}}{2} - \frac{\left| d_{ij} - d_{i'j} \right|}{2} = \max(d_{ij}, d_{i'j}) \tag{15}$$

This algorithm works well when the plotted clusters form distinct clumps (not elongated chains). Application of the procedure presented above leads to well separated, small compact spherical clusters.

3. *Average linkage* - the distance between two clusters is the average distance between all inter-cluster pairs. There are two possible ways of calculating average linkage algorithm: non weighted and weighted, according to the size of each group being compared (*n*). When the clusters' size is equal both algorithms give identical results.

$$d_{mj} = \frac{d_{ij} + d_{i'j}}{2} \text{ (non-weighted)} \tag{16}$$

$$d_{mj} = \frac{n_i}{n} d_{ij} + \frac{n_{i'}}{n} d_{i'j} \quad \text{with} \quad n = n_i + n_{i'} \text{ (weighted)} \tag{17}$$

Applying weighted average linkage algorithm causes no deformation of the clusters. To some extent small clusters consisting of outliers might arise.

4. *Ward's method* is a minimum distance hierarchical method which calculates the sum of squared Euclidean distances from each case in a cluster to the mean of all variables. The cluster to be merged is the one which will increase the sum the least. Thus, this method minimizes the sum of squares of any pair of clusters to be formed at a given step.

$$d_{mj} = \frac{n_i + n_j}{n + n_j} d_{ij} + \frac{n_{i'} + n_j}{n + n_j} d_{i'j} - \frac{n_j}{n + n_j} d_{ii'} \tag{18}$$

5. *Centroid linkage* is calculated as the average of a cluster is applied as the basis for aggregation without distorting the cluster space.

$$d_{mj} = \frac{n_{ij}}{n} d_{ij} + \frac{n_{i'j}}{n} d_{i'j} - \frac{n_i n_{i'}}{n} d_i d_{i'} \tag{19}$$

6. *Median linkage* is calculated as the median of a cluster is applied as the basis for aggregation without distorting the cluster space.

$$d_{mj} = \frac{d_{ij}}{2} + \frac{d_{i'j}}{2} - \frac{d_{ii'}}{4} \tag{20}$$

An advantage of median linkage algorithm lies in preserving the importance of a small cluster after aggregation with a large one.

There are numerous additional linkage algorithms (*correlation of items, binary matching*, etc.), but it would be rare that a researcher needs to apply too many combination

of distance and linkage measures, however comparing of many approaches may be a way of clustering pattern validation.

In hierarchical agglomerative clustering the graphical output of the analysis is usually a dendrogram - a tree-like graphics, which indicates the linkage between the clustered objects with respect to their similarity (distance measure). Decision about the number of statistically significant clusters could be made for different reasons. Often a fixed number of clusters is to be assumed. Sometimes a distance measure or an allowed difference between clusters (classes) is used for evaluating the number of significant clusters. For practical reasons the Sneath's index of cluster significance is widely used. It represents this significance on two levels of distance measure D/D_{max} relation: $1/3D_{max}$ and $2/3D_{max}$. Only clusters remaining compact after breaking the linkage at these two distances are considered significant and are object of interpretation.

The algorithms for non-hierarchical clustering offer the division of the studied objects into *a priori* given number of clusters (determined by some practical or theoretical reasons).

In principle, the data set could be considered as a matrix consisting of rows (the objects) and columns (the features describing the objects). CA makes it possible to classify both the objects and variables. This is very important from practical point of view because in environmetric studies it is very interesting to get information on relationships between the sampling locations and between the monitoring parameters. The whole idea of risk assessment is based actually on estimation of relationship between monitoring features (variables).

Specific goal of the cluster analysis is not only to classify data in a certain way but to detect and visualize the reasons for grouping, e.g. to detect those features which are responsible for the organization of the objects into different structures. With the application of CA an "optimization" of the data is achieved before starting a more detailed data analysis. This way of clustering finds groups of objects or features spontaneously without need of a preliminary training set or preliminary information about the system in consideration.

Let assume that we have n objects, which have to be clustered by the values of m features (variables). The input data matrix X $[n \times m]$ is:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & x_{ij} & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}$$

Each object could be presented by the vector X_i, called object vector i. The final aim is grouping of all n objects according to their features.

As already mentioned above the calculation of a similarity measure is a substantial part of the clustering algorithm. The graphical presentation of the determination of a distance of similarity between two objects O_1 and O_2 located in the space of two features $\boldsymbol{x}$ and $\boldsymbol{y}$ could be found in Figure 1 (Euclidean distance). This distance could be calculated using the Pythagoras theorem

$$d(O_1, O_2) = \sqrt{(y_1 - y_2)^2 + (x_1 - x_2)^2} \tag{21}$$

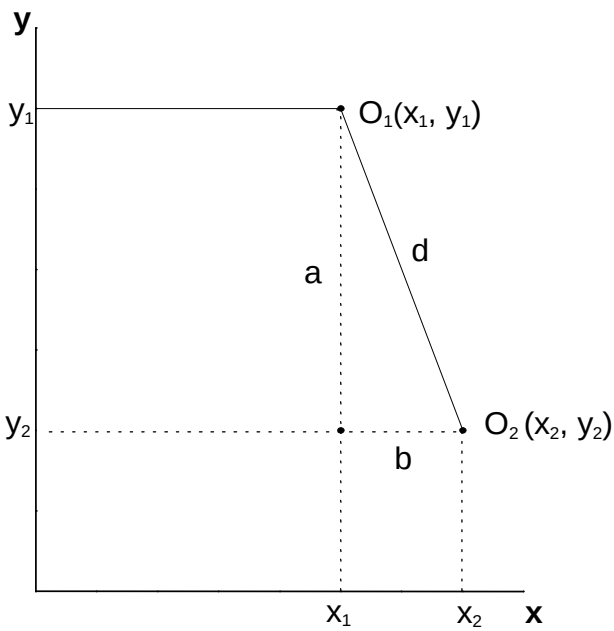

Fig. 1. Euclidean distance between two objects

d(O_i, O_k) is the Euclidean distance and could be easily summarized for m features for each couple of objects O_i and O_k:

$$d(i,k) = \sqrt{\sum_{j=1}^{m} \left(x_{ij} - x_{kj} \right)^2} \tag{22}$$

This is only a partial case from the more general Minkowski distance shown above:

$$d(i,k) = \sqrt[C]{\sum_{j=1}^{m} \left| x_{ij} - x_{kj} \right|^C} \tag{23}$$

where C is a special parameter with values:

- C = 1 - the absolute distance between the objects $d(1,2) = a + b$,
- C = 2 - Euclidean distance,
- C > 2 - in this situation objects - outliers could be determined.

The choice of similarity measure is an obligation and privilege for the chemometrician In most of the situations the Euclidean distances detect better differences between the objects, since the correlation distances are preferred for finding objects similarities.

The determination of the similarity requires transformation of the raw input data due to:

- features are measured in different scales,
- features have different dimensions (orders of magnitude),
- features have different variation.

Unwanted declination of the data structure are avoided by the use of different transformations and scaling procedures and the most sound of them is so called autoscaling (*z*-transform):

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j} \tag{24}$$

where $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and $s_j = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2$. The values z_{ij} are dimensionless and are normally distributed with mean value 0 and variation 1.

From z_{ij} values the Euclidean distances are calculated forming the similarity matrix *D*. This ($n \times n$) matrix is symmetrical with zeros along the main diagonal:

$$D = \begin{pmatrix} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ d_{21} & 0 & d_{23} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & d_{n3} & \dots & 0 \end{pmatrix}$$

After determination of the distances of similarity, the linkage algorithm has to be chosen and one of the most representative graphical plots for the hierarchical agglomerative linkage is the dendrogram.

In most of the environmental studies the agglomerative hierarchical methods are widely used. However, the different linkage algorithms could produce different classification results. Since the correct interpretation of the clustering is the main goal of the studies, it is important to have comparison of the mostly used linkage algorithms.

The scheme offered by Lance and Williams describes comparatively the different linkage options. Let us assume that between the objects i_1 and i_2 the highest level of similarity exists. The distances between the newly formed cluster i_{12} and the rest of the groups (clusters) could be expressed in the following way:

$$d\left(i_{12}, k\right) = \alpha_1 d_{i1,k} + \alpha_2 d_{i2,k} + \beta d_{i1,i2} + \gamma \left| d_{i1,k} - d_{i2,k} \right| \tag{25}$$

where: α_1 - the relative weight of the distance between i_1 and a certain object (cluster), α_2 - the relative weight of the distance between i_2 and a certain object (cluster), β - the relative weight of the distance between i_1 and i_2, γ - the relative weight of the distance between i_1 and i_2 and a certain object (cluster).

In Table 1 the most applied linkage algorithms are presented by combination of the parameters α, β, and γ.

Table 1

Linkage algorithms in cluster analysis

Algorithm	α_1	α_2	β	γ	**Distance**	**Result**
Average linkage	0.5	0.5	0	0	$\frac{1}{2}\left(d_{i1,k} + d_{i2,k}\right)$	Average clusters
Single linkage	0.5	0.5	0	–0.5	$\min\left(d_{i1,k} + d_{i2,k}\right)$	Big clusters
Complete linkage	0.5	0.5	0	0.5	$\max\left(d_{i1,k} + d_{i2,k}\right)$	Small clusters
Ward's method	$f_1(n^*)$	$f_1(n^*)$	$f_1(n^*)$	0	depending on n^*	Balanced clusters

From all these algorithms only the method of Ward uses "*a posteriori*" criterion. Using the heterogeneity of the system expressed by $f(n^*)$, one can decide if a certain object joins a given cluster or not.

The use of fuzzy logic in cluster analysis has created many hierarchical and non-hierarchical algorithms called fuzzy clustering.

Let us assume that a set of objects E with classification structure (C_1, C_2, … C_K) is available. The standard approaches under the convention "hard clustering" suppose that each object of E belongs exactly to a certain cluster C_K. However, the practical application of cluster analysis often leads to another output. A certain object could be associated to greater or lower extent to several clusters simultaneously. That is why in fuzzy clustering the measure t_{ik} is introduced, which presents the level of association of object i to cluster C_K and can accept all possible values within the interval [0,1]:

$$0 \leq t_{ik} \leq 1$$

Thus, for each object that does not belong entirely to one single cluster several values of t_{ik} bigger than zero could be found and for them the following condition is fulfilled:

$$\sum_{i=1}^{K} t_{ik} = 1 \text{ where: } i = 1,2,\ldots,n \tag{26}$$

The definition of the membership of each object to each cluster is given by a vector consisting of t_{ik} values:

$$\vec{t_i} = (t_{i1}, t_{i2}, \ldots, t_{ik}) \tag{27}$$

Fuzzy clustering aims the determination of the membership function for each object and, in such a way, achieving a more reliable "soft clustering" of the object, giving different options for belonging to one or another cluster.

The classification of the objects or the features of a system using cluster analysis could solve the following problems:

- detection of data structure,
- detection of factors responsible for this structure,
- which of the objects are accordingly grouped and could participate in creation in intergroup model,
- which features adequately describe the objects and phenomena studied,
- projection of the multivariate data on plane and construction of respective graphical image.

Principal components regression

In case of many studies related to natural ecosystems PCA and other multivariate statistical techniques are used to determine possible natural or anthropogenic influences in the formation of the determinants total mass. However, PCA does not provide a direct balancing and apportionment. After the pollution sources identification by the application of PCA, the next calculation step in modelling and balancing of pollution impacts is the apportioning itself. It is performed mostly by absolute principal components analysis (APCA). The procedure introduced by Thurston and Spengler is well developed and often applied for apportionment purposes, mainly in apportionment of airborne particulate matter. However, recent applications of the approach proved its effectiveness in apportionment

monitoring studies for other environmental compartments like surface water, soils, sediments, and biota.

The first step in the source apportionment methodology of Thurston and Spengler [8] performing of principal components analysis. As already described the input monitoring data are transformed into a dimensionless standardized form, mostly by z-transform. It is important to note at the outset that, for multivariate analyses such as proposed, sufficient degree of freedom should be available in the model. Thus, the dataset employed must have many more observations m (cases, objects) than variables n (features, parameters). If stable results are to be derived an empirical rule recommends $m \geq n + 50$.

The PCA assumes that the total concentration of each element is made up of the sum of elemental contributions from each of f pollution source components. Hence

$$Z_{ik} = \sum W_{ij} P_{jk} \quad (\text{for } j = 1 \ldots p) \tag{28}$$

where P_{jk} is the jth component's value for observation k; $j = 1,\ldots,p$ is the number of pollution sources influencing the data and W_{ij} is the coefficient matrix of the components.

This equation may be inverted, yielding (in matrix terms)

$$[P]_{jxk} = [B]_{jxi} [Z]_{ixk} \tag{29}$$

where

$$[B]_{jxi} = [W]_{jxi}/\lambda_j \tag{30}$$

where λ_j is the eigenvalue associated with P_j.

The PC scoring matrix [B] is derived so that the first principal component PC1 explains as large a per cent of the original variables' total variance as possible. The coefficients for the second principal component PC2 are, in turn, chosen so that it explains as large a per cent of the remaining variance in the original variables (i.e. not explained by PC1), subject to the restriction than PC1 and PC2 are uncorrelated. In general, the coefficients for $PC_{(j)}$ are chosen so that $PC_{(j)}$ explains as much of the remaining variance (i.e. not explained by PC1 – $PC_{(j-1)}$), subject to the constraint that $PC_{(j)}$ be uncorrelated with the previous PCs.

The PC equation coefficients [B] are mathematically derived from the correlation matrix

$$[R]_{ixi} = [Z]_{ixi} [Z]^t_{ixi} \tag{31}$$

Since the objective of PCA is to find orthogonal (uncorrelated) components, the correlation matrix [R] is diagonalized. The diagonalization finds a matrix Q such that

$$[Q^{-1}]_{ixi} [R]_{ixi} [Q]_{ixi} = [\Lambda]_{ixi} \tag{32}$$

where [Λ] is a diagonal matrix (i.e. a matrix with no off diagonal cross-correlation terms) of eigenvalues arranged in descending order of magnitude and Q contains the corresponding eigenvectors which diagonalize the correlation matrix. By definition, these eigenvectors are the matrix [B], which can be used to derive the PC score matrix [P] from the [Z] matrix.

The primary objective of applying PCA is to derive a small number of components which explain a maximum of the variance in the data. Initially, the PCA results in as many PCs as there are original variables n. Usually, however, only a limited number of these uncorrelated PCs (e.g. five or six) are required to explain virtually all of the variance in a data set of fifteen or more original (intercorrelated) variables. In order for this reduction in the dimensionality to be useful, the new variables (components, latent factors)

must have simple substantive interpretations. Empirically, it has been found that unrotated PCs are often not readily interpretable since they each attempt to explain all remaining variance in the data set. This calculation results in a number of sources of variance being grouped together. For this reason, a limited number of components ($p < n$) are usually subjected to rotation using a criteria such as varimax. After PCA rotation, the resulting components have been found to often be more representative of individual underlying sources of variation. This, in turn, results in more interpretable and useful PCs corresponding, for instance, to different sources of pollution like oil combustion, vehicle emissions, soil dust etc.

The first step in the derivation of source impacts is to calculate component scores for each sample (object). Rotated PC coefficients, B^*, are calculated by applying the rotation transformation matrix [T] to [B]

$$[B]^*_{pxn} = [B]_{pxn}\,[T]_{nxn} \tag{33}$$

Rotated PC scores are computed using the transformed [B] matrix

$$[P]^*_{pxm} = [B]^*_{pxn}\,[Z]_{nxm} \tag{34}$$

These PC scores are correlated with their respective pollution source impacting the site (i.e. a higher component score P^*_{jk} implies a higher pollution impact by the pollution source j during observation k). However, because they are computed from the normalized elemental concentrations Z_{ik}, they too are normalized. Each component indicates deviations from the mean source impacts; they are not proportional to these pollution impacts.

It has been shown that the regression of a dependent variable Y_k on the daily scores of components P_{jk} could be presented by the formula

$$Y_k = Y_a + \sum \zeta_j P_{jk} \quad (\text{for } j = 1\ldots p) \tag{35}$$

where Y_a equals the mean of Y_k. If the dependent variable Y_k is the total mass (for air particulate matter, in [μg m^{-3}]), then ζ_j are the conversion coefficients of the non-dimensional PC score deviations into mass deviations from the mean source impact. Since the components are not scored as deviations from zero, but instead as deviations from the mean, this results in the presence of Y_a in the equation. This, in turn, prevents a direct apportionment of the total particle mass at the sampling site to the normalized pollution source components.

Different studies have addressed the problem of deriving components related to absolute zero by initializing a correlation about the origin (instead of about the mean) and normalizing the data to the mean of all elements in each sample (Q analysis) instead of to the mean of all samples for each element (R analysis). This allowed retention of the information regarding the relative size of each element in absolute terms (i.e. distance from zero). The resultant vectors ware then target transformed to align (to a least square fit) with known pollution source element profile vectors. Another approach attempted to target factors to assumed single element trace compositions, but the data employed were very limited. These techniques of target transformation factor analysis (TTFA) are not readily executed with conventional statistical packages, which do not provide for target rotations.

The already classical approach of Thurston and Spengler offers a simple PCA pollution apportionment procedure which can be executed on conventional, generally available statistical packages such as STATISTICA. The PCA is conducted using a conventional R analysis of elemental concentrations about their means.

As the factor scores obtained from PCA are normalized, with mean zero and standard deviation equal to unity, the true zero for each factor score is calculated by introducing an artificial sample with concentration equal to zero for all variables.

$$(Z_0)_i = \frac{(0 - C_i)}{s_i} = -\frac{C_i}{s_i} \tag{36}$$

where: C_i - arithmetic mean concentration of analyte i (understood as feature), s_i - standard deviation of variable i.

Then the rotated absolute zero PC scores, P^*_0 for each of p components are calculated

$$P^*_{0p} = \sum B^*_{pi}(Z_0)_I \text{ (for } i = 1 \ldots n) \tag{37}$$

These estimates of the PC scores for each component at absolute zero are then used to estimate Absolute PC Scores [*APCS*] for each component on each sampling day as follows:

$$[APCS]^*_{pxj} = [P]^*_{pxj} - [P_0]^*_{pxj} \tag{38}$$

where the j columns of $[P_0]^*$ are all identically equal to the values calculated for $P_{0p}{}^*$. It can be proved in a straight forward manner that the calculation for $[APCS]^*$ gives the exact score which would be achieved had the original scoring been executed using unnormalized data.

Regressing (multiple linear regression) the monitoring results on these *APCS* give estimates of the coefficients which convert the *APCS* into pollutant source mass contributions [μg m^{-3}] from each source for each sample.

The source contributions to C_i can be calculated by mentioned above linear regression procedure according to the following:

$$C_i = (b_0)_i + \sum APCS_p \cdot b_{pi}, \quad p = 1, 2, ..., n \tag{39}$$

where: $(b_0)_i$ - constant term of multiple regression for variable i, b_{pi} - the coefficient of multiple regression of the source p for variable i, $APCS_p$ - scaled value of the rotated factor p for the considered sample, $APCS_p \cdot b_{pi}$ represents the contribution of source p to C_i.

The mean of the product $APCS_p \cdot b_{pi}$ on all samples represents the average contribution of the sources. The method estimates source profiles and contributions but its serious disadvantage is error propagation in centering and uncentering of data. This balancing approach accepts that all sources have been identified by the principal components analysis and all of them participate in the source contribution procedure.

N-way principal components analysis

Most common chemometric techniques are performed using two-way data sets, often presented in the form of matrices. However, due to increasing interest of handling of three-way data matrices three-dimensional analogies to two-dimensional techniques are required. For example, a two-dimensional PCA has its analogue in the form of Tucker3 model [9]. The generality of the Tucker3 model, had made it an often used model for decomposition, compression and interpretation in many applications. The use of multiway models provides a better insight into the data structure, reduces the noise, and shows which of the original variables are correlated and which of them are most significant for a certain environmental problem description. It should be emphasized at this point that the most typical environmental data sets have the following form of a three-dimensional (three-mode) matrix: sampling sites *x* variables *x* time, but examples in which two separate

modes of the data array are formed by the time dimension are also widely known (e.g. months, years). Tucker3 model involves calculating weight matrices corresponding to each of the three modes and is one of the most basic-multi way models used in environmetrics. The model is defined by the decomposition of a three-way matrix *X* into a three-way *core matrix Z* and three two-way loading matrices *A*, *B*, *C* (one for each mode):

$$x_{ijk} \approx \sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{n=1}^{N} a_{il} b_{jm} c_{kn} z_{lmn} + e_{ijk} \tag{40}$$

where *eijk* represents the residual error term (graphical example of decomposition realized by Tucker3 model is presented in Figure 2).

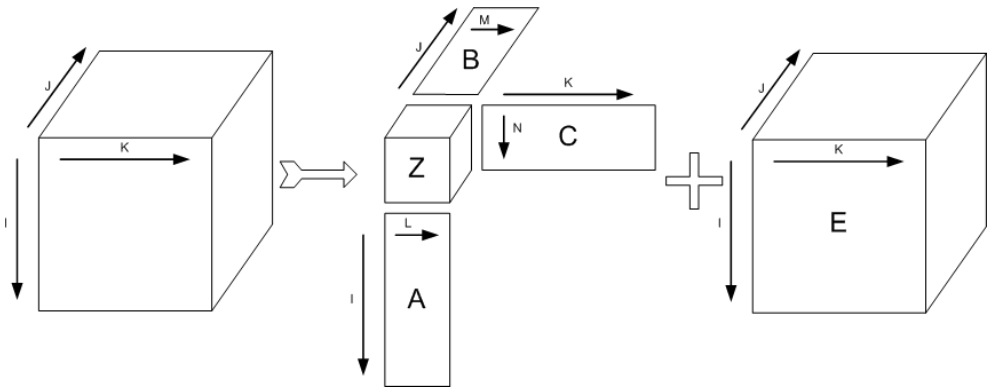

Fig. 2. Tucker3 decomposition model

Tucker3 algorithm delivers a set off possible solutions which mean large number combination of possible models with different complexities. To select a model with an optimal complexity, the variance of each combination of model complexities starting from the model with the lowest number of factor in each mode, to model with the highest complexity should be evaluated. Most often, the optimal model is the one with possibly the smallest number of factors in each of the modes but explaining a large part of data variance. In practice, a trade-off between both requirements is needed. At the same time, a set of possible Tucker3 models should be validated, e.g. using cross-validation procedure. The cross-validation is performed in such a way that part of the data are set to missing, the models are fitted to the remaining data, and the residuals between fitted and true left out elements are calculated.

Self-organizing maps (SOM)

The Self-organizing map (SOM) algorithm has been proposed by Kohonen [10] in 1980. It is a neural-network based model which shares, with the conventional ordination methods, the basic idea of displaying a high-dimensional signal manifold onto a much lower dimensional network in an orderly fashion (usually 2D space). The SOM is a competitive learning algorithm based on unsupervised learning process. This advantage causes no researcher intervention is required during the learning process and that little needs to be known about the characteristics of the input data. In the SOM algorithm, the topological relations and the number of neurons (nodes) organized on a regular

low-dimensional grid are fixed from the very beginning. The number of neurons may vary from a few dozen up to several thousand. Because for SOM algorithm there are no precise rules for the choice of the various, primary defined parameters, the most common shape of the Kohonen map is a rectangular grid with the number of hexagonal nodes (n) determined using following equation:

$$n = 5 \cdot \sqrt{number\ of\ cases} \tag{41}$$

Basically, the two largest eigenvalues of the training data are calculated and the ratio between side lengths of the map grid is set to the ratio between the two maximum eigenvalues. The actual side lengths are then set so that their product is close to the determined number of map units as stated above. Hexagonal lattice is preferred because it does not favor horizontal or vertical direction. Each neutron i is represented by a d-dimensional weight vector (called also prototype vector or codebook vector) $m = [m_i,\ldots,m_d]$, where d is equal to the dimension of the input vectors. The neurons are connected to adjacent neurons by a neighborhood relation, which dictates the topology, or structure of the Kohonen map and, thus, similar objects should be mapped close together on the grid. After primary initialization, the weight vectors are characterized by random values and then SOM is trained iteratively with one of two possible algorithms: sequential or batch. A sequential training algorithm (STA) constructs the nodes in a SOM in order to represent the whole data set and their weights are optimized at each iteration step. In each step, one sample vector x from the input data set is chosen randomly and the distances between it and all the weight vectors of the SOM are calculated using some distance measure, e.g. Euclidean distance, squared Euclidean distance, Mahalanobis distance etc., thus, the optimal topology is expected. Node c, whose weight vector (m) is closest to the input vector x is called the best matching unit (BMU):

$$\|x - m_c\| = \min_i \left\{\|x - m_i\|\right\} \tag{42}$$

where $\|\ \ \|$ is a distance measure (typically Euclidean distance). If missing data appears, they are handled by simply excluding them from the distance calculation (e.g. it is assumed that their contribution to the distance $\|x - m_i\|$ is zero). Because, the same missing value is ignored in each distance calculation (over which the minimum is taken), this is a valid solution. After finding the BMU, the weight vectors are updated in agreement with presented below update rule, so that the BMU is moved closer to the input vector. The topological neighbors of the BMU are moved closer too because of their mutual connection.

$$m_i(t+1) = m_i(t) + \alpha(t) h_{ci}(t)[x(t) - m_i(t)] \tag{43}$$

where: t - time, $m(t)$ - weight vector indicating the output unit's location in the space at time t, $\alpha(t)$ - learning rate at time t, h_{ci} - neighborhood non-increasing function centered in the winner unit c at the time t, $x(t)$ - input vector randomly drawn from the input data set at time t.

The sequential training algorithm is usually performed in two phases. In the first phase, relatively large initial learning rate $\alpha(t = 0)$ and neighborhood radius σ_0 are used. In the second phase both learning rate and neighborhood radius become smaller.

The second possible training algorithm is called batch training algorithm (BTA), because instead of using a single data vector at a time, the whole data set is presented to the map before any adjustments are made. In each training step, the data set is partitioned to the Voronoi regions of the map weight vectors (e.g. each data vector belongs to the data set of the map unit to which it is closest). After that, the new weight vectors are calculated as shown in the following equation:

$$m_i(t+1) = \frac{\sum_{j=1}^{n} h_{ic}(t) x_j}{\sum_{j=1}^{n} h_{ic}(t)} \tag{44}$$

where $c = \arg\min_k \left\{ \left\| x_j \, m_k \right\| \right\}$ is the index of the BMU of data sample x_j.

The new weight vector is a weighted average of the data samples, where the weight of each data sample is the neighborhood function value $h_{ic}(t)$ at its BMU c. Analogous to sequential training algorithm missing values are ignored in calculating the weighted average. The quality of mapping can be quantitatively measured with the quantization error (QE) and the topographic error (TE). After competitive learning process, SOM algorithm enables graphical presentation of results in the form of set of maps (features' planes) and accomplishment of clustering tasks as well. Features' planes can be considered as a sliced version of the SOM map and provide a powerful tool to analyze the community structure. When a plenty of features is considered it is difficult to compare all maps for all features and thus becomes necessary to find similarity between them, and simultaneously, in the cases' space and classify them into clusters. Input features' planes (e.g. variables) could be visualized on a summary SOM map (called also as unified distance matrix or U-matrix) to show the contribution of each feature in the self-organization of the map. U-matrix visualizes distances between neighboring map units, and helps to identify cluster structure of the map: high values of the U-matrix indicates a cluster border, uniform areas of low values indicate clusters themselves, while each feature's plane shows the values of one feature in each map unit. In other words U-matrix expresses semi-quantitative information about the distribution of a complete set of features for a complete set of the cases while separate feature's plane visualizes distribution of a given feature for a complete set of the cases. Because of this, U-matrix joined with features' planes can be effectively applied for assessment of inter-features and inter-cases relations. According to clustering task, one of the most commonly applied algorithm is non-hierarchical K-means clustering algorithm. In this case, different values of k (predefined number of clusters) are tried and the sum of squares for each run is calculated. Finally, the best classification with the lowest Davies-Bouldin index should be chosen (it is a function of the ratio of the sum of within-cluster scatter and between-cluster separation).

The SOM is an algorithm used to visualize and interpret large high-dimensional data sets SOM is unsupervised pattern cognition method similarly to cluster analysis. The main advantage of SOM is the simultaneous classification of variables and objects (sampling locations). Typical applications are visualization of process states or financial results by representing the central dependencies within the data on the map. The map consists of a regular grid of processing units.

A model of some multidimensional observation, eventually a vector consisting of features (variables), is associated with each unit. The map attempts to represent all available observations with optimal accuracy using a restricted set of models. At the same time the

models become ordered on the grid so that similar models are close to each other and dissimilar models far from each other. Fitting of the model vectors is usually carried out by a sequential regression process, where t = 1, 2,... is the step index: For each sample $x(t)$, first the winner index c (best matching unit - BMU, Fig. 3) is identified by the already mentioned condition:

$$\forall i, \left\| x(t) - m_c(t) \right\| \le \left\| x(t) - m_i(t) \right\| \tag{45}$$

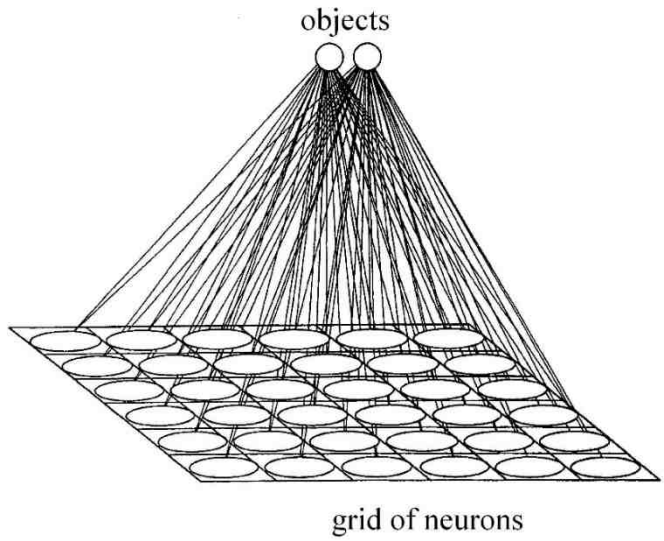

Fig. 3. Self-organizing map architecture

After finding the BMU, the weight vectors of the SOM are updated so that the BMU is moved closer to the input vector in the input space (Fig. 4).

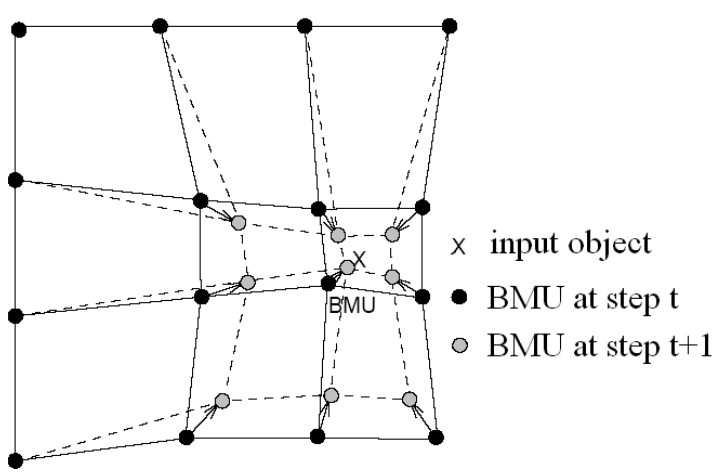

Fig. 4. Updating the BMU and its neighbors towards the input object

Then, all model vectors or a subset of them that belong to nodes centered on node $c = c(\mathbf{x})$ are updated as:

$$m_i(t+1) = m_i(t) + h_{c(x),i}(x(t) - m_i(t)) \tag{46}$$

Here $h_{c(x),i}$ is the "neighborhood function", a decreasing function of the distance between the ith and cth nodes on the map grid. This regression is usually reiterated over the available objects.

The trained map could be graphically presented by 2D planes for each variable indicating variable distribution values on the different map regions by different colors. Additionally the node "coordinates" (vectors) could be clustered by non-hierarchical K-means classification algorithm. Main advantages of SOM algorithm application are:

(i) the projection of variables similarity in the form of features' planes delivers semi-quantitative information about the distribution of a given feature in the space of the cases,
(ii) SOM visualization enables presentation both similarity between positive as well as negative correlated features,
(iii) SOM visualization and SOM-supported classification is able to indicate "outliers" e.g. those features or cases which do not belong to a well-organized, homogeneous populations,
(iv) SOM is noise tolerant (this property is highly desirable when site-measured data are used).

Partial least squares regression

Partial least squares (PLS) regression [11] is often presented as the major regression technique for multivariate data. In fact its use is not always justified by the data, and the originators of the method were well aware of this, but in some applications PLS has been spectacularly successful. In spectroscopy or chromatography we usually expect linear additivity, and this is especially important for chemical instrumental data, and under such circumstances simpler methods such as multiple linear regression (MLR) are often useful, especially when the system in consideration is well known. However, PLS is always an important tool when there is partial knowledge of the data, e.g. NIR spectroscopic measurements of wheat protein. A model can be obtained from a series of wheat samples, and PLS will use typical features in this data set to establish a relationship to the known amount of protein. PLS models can be very robust provided that future samples contain similar features to the original data, but the predictions are essentially statistical.

The most wide spread approach is often called PLS1. Although there are several algorithms, the main ones due to Wold and Martens, the overall principles are quite straightforward. Instead of modelling exclusively the x variables, two sets of models are obtained:

$$\boldsymbol{X} = \boldsymbol{T} \cdot \boldsymbol{P} + \boldsymbol{E} \tag{47}$$

$$\boldsymbol{c} = \boldsymbol{T} \cdot \boldsymbol{q} + \boldsymbol{f} \tag{48}$$

where $\boldsymbol{q}$ has analogies to a loadings vector, although is not normalized. The product of $\boldsymbol{T}$ and $\boldsymbol{P}$ approximates to specific data (e.g. spectral data) and the product of $\boldsymbol{T}$ and $\boldsymbol{q}$ to the true outputs (e.g. concentrations); the common link is $\boldsymbol{T}$. An important feature of PLS is that it is possible to obtain a scores matrix that is common to both the concentrations (c) and measurements (x). It has to be stressed that $\boldsymbol{T}$ and $\boldsymbol{P}$ for PLS are different to $\boldsymbol{T}$ and $\boldsymbol{P}$

obtained in PCA, and unique sets of scores and loadings are obtained for each compound in the data set. Hence if there are 10 compounds of interest, there will be 10 sets of $\boldsymbol{T}$, $\boldsymbol{P}$ and $\boldsymbol{q}$. In this way PLS differs from PCR in which there is only one set of $\boldsymbol{T}$ and $\boldsymbol{P}$.

For an imaginary data set consisting of 25 spectra observed at 27 wavelengths, for which 8 PLS components are calculated, there will be:

- a $\boldsymbol{T}$ matrix of dimensions 25 x 8,
- a $\boldsymbol{P}$ matrix of dimensions 8 x 27,
- an $\boldsymbol{E}$ matrix of dimensions 25 x 27,
- a $\boldsymbol{q}$ vector of dimensions 8 x 1,
- an $\boldsymbol{f}$ vector of dimensions 25 x 1.

Each successive PLS component approximates both the concentration and spectral data better. For each PLS component, there will be a:

- scores vector $\boldsymbol{t}$,
- spectral loadings vector $\boldsymbol{p}$,
- concentration loadings scalar q.

In most of PLS implementations it is conventional to center both the x and c data by subtracting the mean of each column before analysis. Not always this procedure is needed.

For a given compound, the remaining percentage error in the x matrix for a PLS components can be expressed in a variety of ways. The predicted measurements simply involve calculating $\boldsymbol{X_c} = \boldsymbol{T} \cdot \boldsymbol{P}$. The only difference is that each compound generates a separate scores matrix, unlike PCR where there is a single scores matrix for all compounds in the system and so there will be a different behavior in the x block residuals according to the compound.

An extension to PLS1 is often called PLS2; the latter allows the use of a concentration matrix $\boldsymbol{C}$ rather than concentration vectors for each individual compound in the system and the algorithm is iterative. The equations for PLS2 differ slightly in that $\boldsymbol{Q}$ is a matrix (not a vector), so that:

$$\boldsymbol{X} = \boldsymbol{T} \cdot \boldsymbol{P} + \boldsymbol{E} \tag{49}$$

$$\boldsymbol{C} = \boldsymbol{T} \cdot \boldsymbol{Q} + \boldsymbol{F} \tag{50}$$

The number of columns in $\boldsymbol{C}$ and $\boldsymbol{Q}$ are equal to the number of objects (e.g. compounds) of interest. In PLS1 one compound is modelled at a time, whereas in PLS2 all known compounds can be simultaneously included in the model.

Conclusion

The presented lecture is dedicated to master students but could be of use for any student interested in data mining approaches. In the references some major textbook are given and each one of them could be source of additional information on the theory and application of chemometrics and environmetrics for data interpretation and modelling. Additional small selection of applications of exploratory data analysis are offered for individual work and seminars [12-17].

References

[1] Simeonov V, Einax J, Stanimirova I, Kraft J. Anal Bioanal Chem. 2002;374:898-905. DOI: 10.1007/s00216-002-1559-5

[2] Tobiszewski M, Nedyalkova M, Madurga S, Pena-Pereira F, Namieśnik J. Ecotoxicol Environ Saf. 2018;147:292-8. DOI: 10.1016/j.ecoenv.2017.08.054.
[3] Yuan J, Chun X, Zhang T, Sun J, Yuan X, Yu S, et al. Chemosphere. 2016;156:334-40. DOI: 10.1016/j.chemosphere.2016.05.002.
[4] Olkowska E, Kudłak B, Tsakovski S, Ruman M, Simeonov V, Polkowska Z. Sci Total Environ. 2014;476:477-84. DOI: 10.1016/j.scitotenv.2014.01.044.
[5] Nedyalkova M, Simeonov V. Open Chem. 2019;17:711-21. DOI: 10.1515/chem-2019-0082.
[6] Einax J, Zwanziger H, Geiss S. Chemomerics in Environmental Analysis. Weinheim: VCH, Wiley; 1997. ISBN: 3527287728.
[7] Brown S, Tauler R, Walczak B. Comprehensive Chemometrics. Amsterdam: Elsevier; 2019. ISBN: 9780444641656.
[8] Thurston G, Spengler J. Atmos Environ. 1985;19:9-25. DOI: 10.1016/0004-6981(85)90132-5.
[9] Kroonenberg P, de Leeuw J. Psychometrica. 1980;45:69-97. DOI: 0033-3123/80/0300-2831500.75/0.
[10] Kohonen T. Self-organizing Maps. Berlin: Springer Science; 1997. ISBN: 9783540679219.
[11] Massart DL, Vandeginste B, Buydens L, de Jong S, Lewi P, Smeyers-Verbeke J. Handbook of Chemometrics and Qualimetrics. Amsterdam: Elsevier; 1998. ISBN: 9780444828538.
[12] Pieszczek L, Daszykowski M. Chemom Intel Lab Syst. 2019;187:28-40. DOI: 10.1016/j.chemolab.2019.02.009.
[13] Hewitt J, Hoeting J, Done J, Towler E. Environmetrics. 2018;29:2523. DOI: 10.1002/env.2523.
[14] Lee J, Sun Y, Chang H. Environmetrics. 2019;online. DOI: 10.1002/env.2578.
[15] Ballesteros-Gomez A, Rubio S. Anal Chem. 2011;83:4579-13. DOI: 10.1021/ac200921j.
[16] Zhao H, Grafstrom A. Environmetrics. 2020;31:2625. DOI: 10.1002/env.2625.
[17] Simeonov V. Chem Didact Ecol Metrol. 2019;24(1-2):7-21. DOI: 10.2478/cdem-2019-0001.